

Машинное обучение. Выпуклые задачи. Регуляризация и стабильность

Алексей Колесов

Белорусский государственный университет

11 октября 2017 г.

План

- рассмотреть класс выпуклых задач
- показать, какие выпуклые задачи допускают (эффективное) решение
- рассмотреть понятия «регуляризация» и «стабильность»

Содержание

1 Выпуклые задачи

- Определения
- Выпуклые задачи машинного обучения
- Изучаемость выпуклых задач
- Суррогатные функции потерь

2 Регуляризация и стабильность

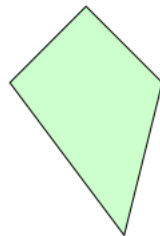
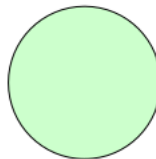
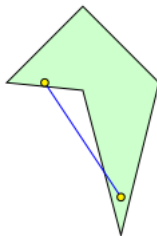
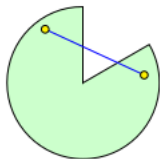
- Минимизация регуляризированной функции потерь
- Стабильные алгоритмы не переобучаются
- Регуляризация Тихонова (L2)
- Fitting-stability tradeoff

Выпуклое множество

Выпуклое множество

Подмножество C векторного пространства называется **выпуклым** (convex), если для любых u, v из C отрезок, соединяющий эти два вектора тоже лежит в C . Т.е. $\forall \alpha \in [0, 1]$ вектор $\alpha u + (1 - \alpha)v \in C$

Примеры



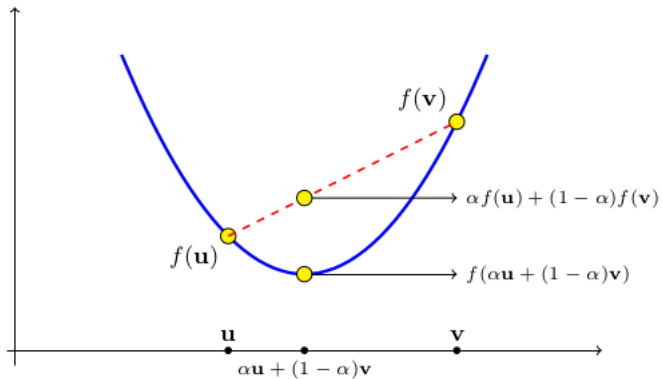
Выпуклая функция

Выпуклая функция

Пусть C — выпуклое множество. Тогда функция f называется **выпуклой** (convex), если для любых двух векторов u, v из C график f лежит под отрезком, соединяющим $f(u)$ и $f(v)$. Т.е, для $\forall u, v \in C$ и $\forall \alpha \in [0, 1]$:

$$f(\alpha u + (1 - \alpha)v) \leq \alpha f(u) + (1 - \alpha)f(v)$$

Пример



Надграфик

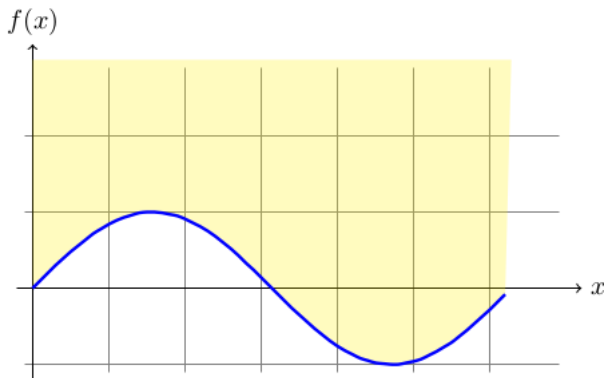
Выпуклая функция

Надграфиком (epigraph) функции называется множество точек, лежащих на или над графиком функции

Лемма о надграфике

Функция f выпукла \iff надграфик f — выпуклое множество

Надграфик



Лемма о минимуме выпуклой функции

Шаром радиуса r и центром в u называют множество точек $B(u, r) = \{v : \|v - u\| \leq r\}$

Точка u называется **локальным минимумом** f , если $\exists r > 0$, что для любого $v \in B(u, r)$ выполняется, что $f(v) \geq f(u)$

Лемма о минимуме выпуклой функции

Любой локальный минимум выпуклой функции является её глобальным минимумом

Доказательство леммы у минимуме выпуклой функции

Имеем: u — локальный минимум выпуклой f Хотим: $\forall v$
 $f(v) \geq f(u)$

Доказательство: Зафиксируем v

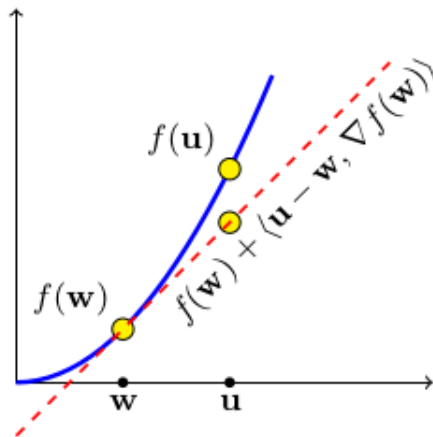
- $\exists \alpha > 0$, такое что $u + \alpha(v - u) \in B(u, r)$
- $f(u) \leq f(u + \alpha(v - u))$
- $f(u + \alpha(v - u)) = f(\alpha v + (1 - \alpha)u) \leq (1 - \alpha)f(u) + \alpha f(v)$
- $\alpha f(u) \leq \alpha f(v)$

Опорные плоскости

- для выпуклой функции f можно в **любой** точке w можно построить касательную плоскость
- (опорная) плоскость будет лежать не выше графика f
- если f дифференцируема, то
$$l(u) = f(w) + \langle \nabla f(w), u - w \rangle$$
 — опорная
- $\nabla f(w) = \left(\frac{\partial f(w)}{\partial w_1}, \dots, \frac{\partial f(w)}{\partial w_d} \right)$

$$\forall u, f(u) \geq f(w) + \langle \nabla f(w), u - w \rangle$$

Пример



Критерий выпуклости

Критерий выпуклости

Пусть $f : \mathbb{R} \rightarrow \mathbb{R}$ дважды дифференцируема. Тогда следующие утверждения эквивалентны:

- f выпукла
 - f' монотонно не убывает
 - $f'' \geq 0$
-
- $f(x) = x^2$ выпукла $\Leftrightarrow f''(x) = 2 > 0$
 - $f(x) = \log(1 + \exp(x))$ выпукла \Leftrightarrow
$$f'(x) = \frac{\exp(x)}{1 + \exp(x)} = \frac{1}{\exp(-x) + 1} \text{ возрастает}$$

Лемма о композиции выпуклой и линейной функции

Лемма о композиции выпуклой и линейной функции

Пусть $f : \mathbb{R}^d \rightarrow \mathbb{R}$ может быть представлена как $f(w) = g(\langle w, x \rangle + y)$, где $x \in \mathbb{R}^d$, $y \in \mathbb{R}$, $g : \mathbb{R} \rightarrow \mathbb{R}$. Тогда g — выпукла $\Rightarrow f$ — выпукла

$$f(\alpha w_1 + (1 - \alpha)w_2) = g(\langle \alpha w_1 + (1 - \alpha)w_2, x \rangle + y) \quad (1)$$

$$= g(\alpha \langle w_1, x \rangle + (1 - \alpha) \langle w_2, x \rangle + y) \quad (2)$$

$$= g(\alpha(\langle w_1, x \rangle + y) + (1 - \alpha)(\langle w_2, x \rangle + y)) \quad (3)$$

$$\leq \alpha g(\langle w_1, x \rangle + y) + (1 - \alpha)g(\langle w_2, x \rangle + y) \quad (4)$$

Примеры

- пусть $x \in \mathbb{R}^d$, $y \in \mathbb{R}$ и $f(w) = (\langle w, x \rangle - y)^2$. Тогда $f(w)$ — выпукла
- пусть $x \in \mathbb{R}^d$, $y \in \{-1; 1\}$, тогда $f(w) = \log(1 + \exp(-y\langle w, x \rangle))$ — выпукла

Лемма о максимуме и сумме выпуклых функций

Лемма о максимуме и сумме выпуклых функций

Пусть для $i = 1, \dots, r$ $f_i : \mathbb{R} \rightarrow \mathbb{R}$ — выпуклые функции. Тогда следующие функции тоже являются выпуклыми:

- $g(x) = \max_{i \in [r]} f_i(x)$
- $\sum_{i=1}^r w_i f_i(x)$, где $w_i \geq 0$

Например, $f(x) = |x|$ выпукла.

Липшицевость

Липшицевость

Пусть $C \subset \mathbb{R}^d$. Функция $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$ называется ρ -липшицева, если для любых w_1, w_2 из C выполняется,

$$\|f(w_1) - f(w_2)\| \leq \rho \|w_1 - w_2\|$$

По теореме о среднем значении $f(w_1) - f(w_2) = f'(u)(w_1 - w_2)$, поэтому если $|f'| < \rho$, то функция ρ -липшицева

Примеры

- $f(x) = |x|$ 1-липшицева \Leftarrow
 $|x_1| - |x_2| = |x_1 - x_2 + x_2| - |x_2| \leq |x_1 - x_2| + |x_2| - |x_2| = |x_1 - x_2|$
- $f(x) = \log(1 + \exp(x))$ 1-липшицева, так как
$$|f'(x)| = \left| \frac{1}{1 + \exp(-x)} \right| \leq 1$$
- $f(x) = x^2$ на \mathbb{R} не липшицева для любого ρ
- $f(x) = x^2$ на $[-a; a]$ липшицева с $\rho = 2a$
- $f(w) = \langle w, v \rangle + b \|v\|$ -липшицева по неравенству Коши

Лемма о композиции липшицевых функций

Лемма о композиции липшицевых функций

Композиция ρ_1 и ρ_2 липшицевых функций является $(\rho_1 \rho_2)$ -липшицевой

Гладкость

Гладкая функция

Дифференцируемая функция $f : \mathbb{R}^d \rightarrow \mathbb{R}$ называется β -гладкой (smooth), если её градиент β -липшицевый, т.е.

$$\forall v, w \text{ выполняется, что } \|\nabla f(w) - \nabla f(v)\| \leq \beta \|w - v\|$$

- для гладкой функции выполняется,
$$f(v) \leq \nabla f(w) + \langle \nabla f(w), v - w \rangle + \frac{\beta}{2} \|v - w\|^2$$
- если функция гладкая и неотрицательная, то
$$\|\nabla f(w)\|^2 \leq 2\beta f(w) \text{ (самоограниченная)}$$

Примеры

- $f(x) = x^2$ является 2-гладкой ($f'(x) = 2x$)
- $f(x) = \log(1 + \exp(x))$ является $(1/4)$ -гладкой (см. $f''(x)$)

Лемма о композиции гладкой и линейной функции

Лемма о композиции гладкой и линейной функции

Пусть $f(w) = g(\langle w, x \rangle + b)$, причём g является β -гладкой.
Тогда f является $(\beta\|x\|^2)$ -гладкой.

- $f(w) = (\langle w, x \rangle + b)^2$ является $(2\|x\|^2)$ -гладкой
- $f(w) = \log(1 + \exp(-y\langle w, x \rangle))$ является $(\|x\|^2/4)$ -гладкой

Определения

Задача выпуклой оптимизации

Задача минимизации выпуклой функции на выпуклом множестве называется **задачей выпуклой оптимизации** (convex optimization problem)

Выпуклая задача машинного обучения

Задача машинного обучения (H, Z, l) называется **выпуклой** (convex learning problem), если множество гипотез H является выпуклым, и l является выпуклой для любого $z \in Z$.

Пример

Рассмотрим линейную регрессию с квадратичной функцией потерь.

Раньше задавали так: $H = \{x \mapsto \langle w, x \rangle, w \in \mathbb{R}^d\}$,

$$l(h, (x, y)) = (h(x) - y)^2$$

Теперь:

- $H = \mathbb{R}^d$
- $Z = X \times Y = \mathbb{R}^{d+1}$
- $l(w, (x, y)) = (\langle w, x \rangle - y)^2$
- задача — выпукла

Лемма о выпуклых задачах машинного обучения

Лемма о выпуклых задачах машинного обучения

Минимизация эмпирического риска для выпуклой задачи машинного обучения является задачей выпуклой оптимизации

$$\text{ERM}_H(S) = \underset{w \in H}{\operatorname{argmin}} L_S(w) = \underset{w \in H}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^m l(w, z_i)$$

Все ли выпуклые задачи изучаемы

- задача полупространств изучаема в \mathbb{R}^d (VC-теория)
- с помощью «discretization trick» любая задача с d параметрами изучаема
- все задачи \mathbb{R}^d изучаемы?

Не все выпуклые задачи изучаемы

Пусть $H = \mathbb{R}$, $l(w, (x, y)) = (wx - y)^2$.

Возьмём $\epsilon = 1/4$, $\delta = 0.01$, $m \geq m(\epsilon, \delta)$, $\mu = \frac{\log(100/99)}{2m}$,
 $z_1 = (1, 0)$, $z_2 = (\mu, -1)$

Распределения:

$$D_1(x) = \begin{cases} \mu & \text{если } x = z_1 \\ 1 - \mu & \text{если } x = z_2 \end{cases}.$$

$$D_2(x) = 1_{x=z_2}$$

- вероятность, что S содержит только z_2 больше 0.99
 $((1 - \mu) \geq e^{-2\mu m} = 0.99)$
- если $A(S) = \bar{w} < -1/(2\mu)$, то
 $L_{D_1}(\bar{w}) - L_{D_1}(0) = \frac{1}{4\mu} - (1 - \mu) > 1/4$
- $\bar{w} \geq -1/(2\mu)$, то $L_{D_2}(\bar{w}) \geq 1/4$

Изучаемые классы

Выпукло-липшицево-ограниченная задача

Задача называется **выпукло-липшицево-ограниченной** (convex-lipschitz-bounded) с параметрами ρ, B , если:

- H является выпуклым множеством и $\forall w \in H, \|w\| \leq B$
- $\forall z \in Z$ функция $I(w, z)$ является выпуклой и ρ -липшицевой

Например, пусть $X = \{x \in \mathbb{R}^d : \|x\| < \rho\}$, $Y = \mathbb{R}$,
 $H = \{w \in \mathbb{R}^d, \|w\| < B\}$ и $I(w, (x, y)) = |\langle w, x \rangle - y|$. Данная задача выпукло-липшицево-ограниченная

Изучаемые классы

Выпукло-гладко-ограниченная задача

Задача называется **выпукло-гладко-ограниченной** (convex-smooth-bounded) с параметрами β, B , если:

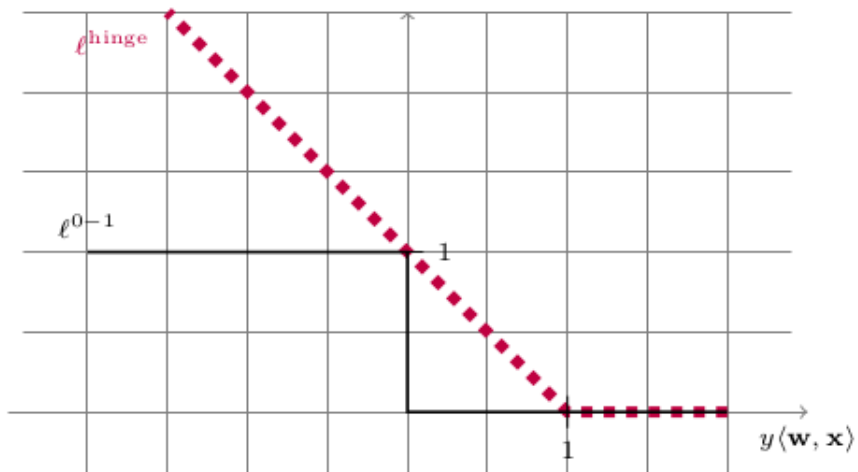
- H является выпуклым множеством и $\forall w \in H, \|w\| \leq B$
- $\forall z \in Z$ функция $I(w, z)$ является выпуклой, неотрицательной и β -гладкой

Например, пусть $X = \{x \in \mathbb{R}^d : \|x\| < \beta/2\}$, $Y = \mathbb{R}$, $H = \{w \in \mathbb{R}^d, \|w\| < B\}$ и $I(w, (x, y)) = (\langle w, x \rangle - y)^2$. Данная задача выпукло-липшицево-ограниченная

Суррогатные функции потерь

- интересные функции потерь часто сложно оптимизировать (0-1 loss)
- можно задать функцию
 - выпукла
 - оценка сверху на оригинальную функцию потерь
- оптимизировать верхнюю границу

Hinge-loss



Декомпозиция

Ошибка такой задачи складывается из:

- approximation error — насколько хорош класс
- estimation error — насколько хорошо решили задачу в неполных данных
- optimization error — насколько велика разница между суррогатной и оригинальной функцией потерь

Содержание

1 Выпуклые задачи

- Определения
- Выпуклые задачи машинного обучения
- Изучаемость выпуклых задач
- Суррогатные функции потерь

2 Регуляризация и стабильность

- Минимизация регуляризированной функции потерь
- Стабильные алгоритмы не переобучаются
- Регуляризация Тихонова (L2)
- Fitting-stability tradeoff

RLM

Будем искать решение вот так:

$$\operatorname{argmin}_w (L_S(w) + R(w))$$

- $R(w)$ может отражать «сложность» гипотезы
- $H = \bigcup_i \{w : R(w) \leq i\}$ (см. SRM)
- $R(w) = \lambda \|w\|^2$ — регуляризация Тихонова
- гребневая регрессия использует регуляризацию
- выбор регуляризации — наложение prior distribution на w

Стабильность

- интуитивно, если «немного» изменить S , то $A(S)$ должен меняться немного (стабильность)
- если $L_D(A(S)) \gg L_S(A(S))$, то есть «переобучение»
- если алгоритм стабилен, то $\mathbb{E}_S[L_D(A(S)) - L_S(A(S))]$ невелико

Определение стабильности

- пусть $S = (z_1, \dots, z_m) \sim D^m$, $z' \sim D$
- $S^{(i)} = (z_1, \dots, z_{i-1}, z', z_{i+1}, z_m)$
- замена S на $S^{(i)}$ — «небольшое изменение входа»
- для «хороших» A $I(A(S^{(i)}), z_i) - I(A(S), z_i) \geq 0$
- если $I(A(S^{(i)}), z_i) - I(A(S), z_i)$ велико, то «переобучение»

Лемма о «небольшом изменении входа»

Лемма о «небольшом изменении входа»

Пусть $U(m)$ — равномерное распределение над $[m]$. Тогда:

$$\begin{aligned} \mathbb{E}_{S \sim D^m} [L_D(A(S)) - L_S(A(S))] = \\ \mathbb{E}_{(S, z') \sim D^{m+1}, i \sim U(m)} [l(A(S^{(i)}), z_i) - l(A(S), z_i)] \end{aligned}$$

- $\mathbb{E}_S [L_D(A(S))] = \mathbb{E}_{S, z_i} [l(A(S), z_i)] = \mathbb{E}_{S, z'} [l(A(S), z_i)]$
- $\mathbb{E}_S [L_S(A(S))] = \mathbb{E}_{S, i} [l(A(S), z_i)]$

В среднем стабильный при замене одного объекта алгоритм

В среднем стабильный при замене одного объекта алгоритм

Алгоритм называется **в среднем стабильным при замене одного объекта** (on-average-replace-on-stable), если существует монотонно убывающая $\epsilon : \mathbb{N} \rightarrow \mathbb{R}$, такая что

$$\mathbb{E}_{(S, z') \sim D^{m+1}, i \sim U(m)} [I(A(S^{(i)}, z_i) - I(A(S), z_i)] \leq \epsilon(m)$$

Стабильность не означает «хорошесть» алгоритма!

План

- стабильные алгоритмы не переобучаются
- докажем, что RLM с регуляризацией Тихонова стабилен
- будем считать, что функция потерь выпукла

Сильная выпуклость

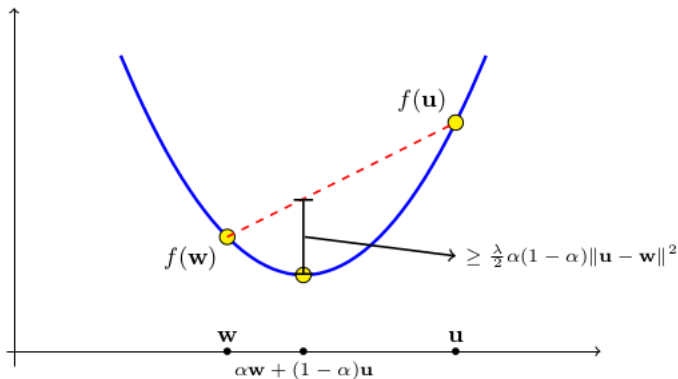
Сильная выпуклость

Функция f называется λ -сильно выпуклой, если для всех u, w и $\alpha \in [0, 1]$ выполняется

$$f(\alpha w + (1 - \alpha)u) \leq \alpha f(w) + (1 - \alpha)f(u) - \frac{\lambda}{2}\alpha(1 - \alpha)\|w - u\|^2$$

Выпуклая функция является 0-сильно выпуклой.

Рисунок



Лемма о сильно выпуклых функциях

Лемма о сильно выпуклых функциях

- ❶ $f(w) = \lambda \|w\|^2$ является 2λ -строго выпуклой
- ❷ если f λ -строго выпуклая, а g — выпуклая, то $f + g$ является λ -строго выпуклой
- ❸ если f является λ -строго выпуклая, и u минимизирует $f(u)$, то для любого w выполняется:

$$f(w) - f(u) \geq \lambda \|w - u\|^2$$

$$\frac{f(u + \alpha(w - u)) - f(u)}{\alpha} \leq f(w) - f(u) - \frac{\lambda}{2}(1 - \alpha)\|w - u\|^2$$

Рассмотрим предел при $\alpha \rightarrow 0$

Регуляризация Тихонова стабилизирует задачу

Рассматриваем:

$$A(S) = \underset{w}{\operatorname{argmin}} (L_S(w) + \lambda \|w\|^2)$$

$f_S(w) = L_S(w) + \lambda \|w\|^2$ является 2λ -строго выпуклой и для любого v

$$f_S(v) - f_S(A(S)) \geq \lambda \|v - A(S)\|^2$$

$$f_S(v) - f_S(u) = L_S(v) + \lambda \|v\|^2 - (L_S(u) + \lambda \|u\|^2) \quad (5)$$

$$= L_{S^{(i)}}(v) + \lambda \|v\|^2 - (L_{S^{(i)}}(u) + \lambda \|u\|^2) \quad (6)$$

$$+ \frac{l(v, z_i) - l(u, z_i)}{m} + \frac{l(v, z') - l(u, z')}{m} \quad (7)$$

Регуляризация Тихонова стабилизирует задачу

Имеем:

$$f_S(v) - f_S(A(S)) \geq \lambda \|v - A(S)\|^2$$

$$f_S(v) - f_S(u) = L_{S^{(i)}}(v) + \lambda \|v\|^2 - (L_{S^{(i)}}(u) + \lambda \|u\|^2) \quad (8)$$

$$\frac{l(v, z_i) - l(u, z_i)}{m} + \frac{l(v, z') - l(u, z')}{m} \quad (9)$$

Возьмем $v = A(S^{(i)})$, $u = A(S)$ и:

$$\begin{aligned} f_S(A(S^{(i)})) - f_S(A(S)) &\leq \frac{l(A(S^{(i)}), z_i) - l(A(S), z_i)}{m} + \frac{l(A(S^{(i)}), z') - l(A(S), z')}{m} \\ \lambda \|A(S^{(i)}) - A(S)\|^2 &\leq \frac{l(A(S^{(i)}), z_i) - l(A(S), z_i)}{m} + \frac{l(A(S^{(i)}), z') - l(A(S), z')}{m} \end{aligned}$$

Случай липшицевой функции потерь

Если l ρ -липшицева, то

$$\begin{aligned}l(A(S^{(i)}), z_i) - l(A(S), z_i) &\leq \rho \|A(S^{(i)}) - A(S)\| \\l(A(S^{(i)}), z') - l(A(S), z') &\leq \rho \|A(S^{(i)}) - A(S)\|\end{aligned}$$

Получаем

$$\begin{aligned}\lambda \|A(S^{(i)}) - A(S)\|^2 &\leq \frac{2\rho \|A(S^{(i)}) - A(S)\|}{m} \\ \|A(S^{(i)}) - A(S)\| &\leq \frac{2\rho}{\lambda m} \\ l(A(S^{(i)}), z_i) - l(A(S), z_i) &\leq \frac{2\rho^2}{\lambda m}\end{aligned}$$

Регуляризация Тихонова стабилизирует RLM

Лемма о RLM в случае выпукло-липшицево-ограниченной задачи

RLM в случае выпукло-липшицево-ограниченной задачи является стабильным с $\epsilon(m) = \frac{2\rho^2}{\lambda m}$. Т.е.

$$\mathbb{E}_{S \sim D^m} [L_D(A(S)) - L_S(A(S))] \leq \frac{2\rho^2}{\lambda m}$$

Чуть более слабую оценку можно доказать для выпукло-гладко-ограниченных задач

Fitting-stability tradeoff

- чем больше λ , тем стабильней алгоритм, но хуже приближение
- $\mathbb{E}_S[L_D(A(S))] = \mathbb{E}_S[L_S(A(S))] + \mathbb{E}_S[L_D(A(S)) - L_S(A(S))]$
- второе слагаемое соответствует стабильности
- хотим, чтоб сумма была маленькой
- можно находить баланс валидацией

Оценки на эмпирический риск

Оценка на эмпирический риск для липшицевой функции потерь

Если мы используем RLM-алгоритм для липшицевой функции потерь с регуляризацией Тихонова, то:

$$\forall w^*, \mathbb{E}_S[L_D(A(S))] \leq L_D(w^*) + \lambda \|w^*\|^2 + \frac{2\rho^2}{\lambda m}$$

Пусть $\lambda = \sqrt{\frac{2\rho^2}{B^2 m}}$, то:

$$\mathbb{E}_S[L_D(A(S))] \leq \min_{w \in H} L_D(w) + \rho B \sqrt{\frac{8}{m}}$$

Содержание

1 Выпуклые задачи

- Определения
- Выпуклые задачи машинного обучения
- Изучаемость выпуклых задач
- Суррогатные функции потерь

2 Регуляризация и стабильность

- Минимизация регуляризированной функции потерь
- Стабильные алгоритмы не переобучаются
- Регуляризация Тихонова (L_2)
- Fitting-stability tradeoff

Итоги

- рассмотрели выпуклость, гладкость, липшицевость функций
- показали, что выпуклости недостаточно для изучаемости
- выделили классы выпуклых задач, которые можно решить
- ввели понятие регуляризации и стабильности
- доказали, что стабильные алгоритмы не переобучаются

Литература

- Shai Shalev-Shwartz and Shai Ben-David — Understanding Machine Learning: From theory to algorithms (главы 12,13)